Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W653–W658***
doi:10.1093/nar/gkm293

COMPASS server for remote homology inference

Ruslan I. Sadreyev[1,*], Ming Tang[1], Bong-Hyun Kim[2] and Nick V. Grishin[1,2]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received January 31, 2007; Revised March 30, 2007; Accepted April 12, 2007

ABSTRACT

COMPASS is a method for homology detection and local alignment construction based on the comparison of multiple sequence alignments (MSAs). The method derives numerical profiles from given MSAs, constructs local profile-profile alignments and analytically estimates E-values for the detected similarities. Until now, COMPASS was only available for download and local installation. Here, we present a new web server featuring the latest version of COMPASS, which provides (i) increased sensitivity and selectivity of homology detection; (ii) longer, more complete alignments; and (iii) faster computational speed. After submission of the query MSA or single sequence, the server performs searches versus a user-specified database. The server includes detailed and intuitive control of the search parameters. A flexible output format, structured similarly to BLAST and PSI-BLAST, provides an easy way to read and analyze the detected profile similarities. Brief help sections are available for all input parameters and output options, along with detailed documentation. To illustrate the value of this tool for protein structure-functional prediction, we present two examples of detecting distant homologs for uncharacterized protein families. Available at http://prodata.swmed.edu/compass

INTRODUCTION

Accurate detection of sequence similarity between distantly related proteins is essential for many fields, including protein structure prediction, protein engineering, and comparative genomics. The performance of an automatic method for sequence comparison can be characterized by sensitivity, selectivity and accuracy of produced sequence alignments. All these parameters can be significantly improved by comparing multiple sequence alignments (MSAs) rather than individual sequences. The improvement comes from evolutionary information about residue preferences at sequence positions in the family represented by the MSA. This information can be extracted from MSAs in two numerical forms: ‘traditional’ position-specific profiles and hidden Markov models (HMMs). The well-known and popular methods for profile-sequence or HMM-sequence comparison include PSI-BLAST (1,2), HMMER (3), SAM-T (4,5) and others. A newer generation of methods involves the comparison of two profiles (6–10) or two HMMs (11,12), with several corresponding web servers available (13–16). These methods further improve the quality of homology detection and alignment construction (17,18). There is a number of publicly available web servers aimed at protein structure prediction that use these and a variety of other techniques [for example, (19–23)].

COMPASS (9) is an established method for profile-based comparison of MSAs. COMPASS derives numerical profiles from given MSAs, constructs optimal local profile-profile alignments, and analytically estimates E-values for the detected similarities. As previously shown by us (9) and independently verified by others (12,18), COMPASS is a sensitive and selective tool for detection of remote sequence similarity that offers accurate local alignments. In many cases, COMPASS provides accurate homology detection and structure prediction that would be difficult or impossible to produce by PSI-BLAST (9,24).

As a standalone package, COMPASS has been used by different research groups (24–31). Until now, COMPASS was only available for download and local installation. Here, we present a new web server featuring the recently improved version of COMPASS.

METHODS

To compare two MSAs, COMPASS performs four steps: (i) processing input MSAs and generating numerical profiles; (ii) calculating scores between individual positions of the compared profiles; (iii) finding optimal local alignment of the two profiles; and (iv) assessing statistical significance of the optimal alignment score (9).

Methodologically, COMPASS is a generalization to profile-profile comparison of the PSI-BLAST approach to

*To whom correspondence should be addressed. Tel: 214-645-5951; Fax: 214-645-5948; Email: sadreyev@chop.swmed.edu profile-sequence comparison. Numerical profiles represent effective counts and frequencies of 21 symbols (20 residue types and gaps) at each position of the input MSAs. To search with a query MSA against a database of MSAs, the database profiles are pre-computed in advance. Scores for the similarity between individual profile positions are calculated using our original formula (9) and then rescaled so that their distribution is similar to a standard distribution with well-known properties (such as BLOSUM62 substitution scores). Rescaled positional scores are used to find the optimal local alignment using the Smith–Waterman algorithm. The statistical significance of the optimal alignment score is estimated using a simple formula for E-value (the expected number of hits in a random database with a score equal to or greater than the observed score). The parameters of this formula are based on our extensive simulations of random profile comparisons (9). As the final result of the search, a list of the most significant hits for the submitted query is displayed, followed by the optimal profile-profile alignments.

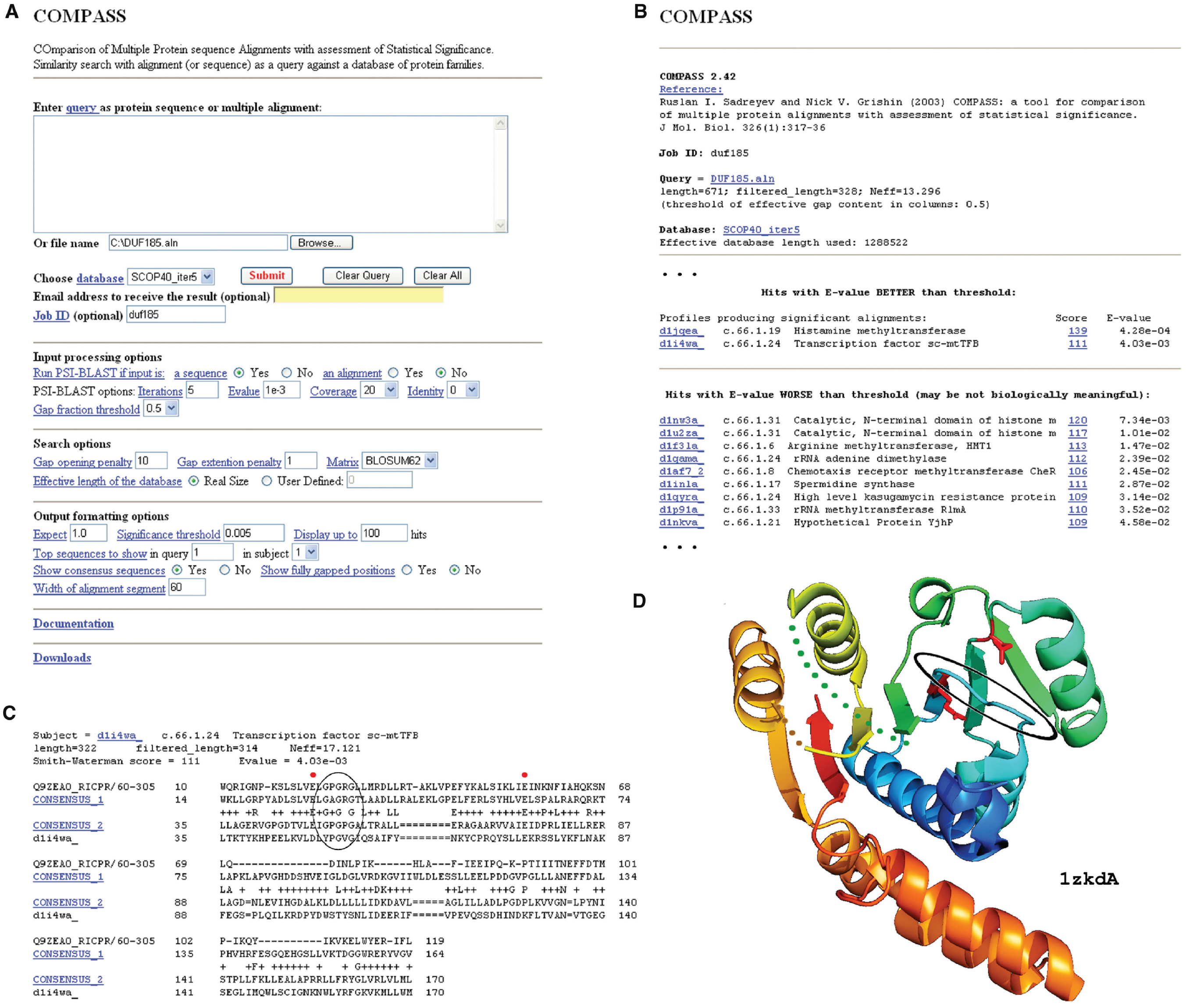
According to our results (9) and independent evaluations (12,18), COMPASS performance has been demonstrated to be among the top methods for profile comparison, by both the quality of homology detection and the accuracy of local alignment construction. The presented web server features a newer version of COMPASS, with several major modifications to improve performance.

(i) *Higher quality of homology detection*. Evaluation of the statistical significance of hits is improved by using a more realistic null model of random profile comparison. The original random model involved the profiles composed of randomly sampled positions from real MSAs. The score statistics were modeled depending on the profile lengths only, and a rough linear approximation of the dependency was used (9). We developed a new random model that captures additional important features of real profiles. First, in order to reproduce local correlations between different positions of MSA, we generate random profiles from fragments of real profiles corresponding to individual elements of secondary structure. Second, to model more accurately the distribution parameters K and λ (2,9) for optimal profile-profile scores, we introduce their dependence on the profile 'thickness' (sequence divergence within the profiles). Finally, we use more precise non-linear functions (combinations of quadratic and square-root) to describe the dependency of these parameters on profile length and 'thickness'. According to our preliminary results, the new version of COMPASS shows roughly 20–25% improvement in the quality of similarity detection.

(ii) *Longer, more complete local alignments*. Rescaling of individual positional scores is modified, so that alignment coverage increases. In the original version, this procedure was similar to the composition-based statistic in PSI-BLAST (2), which standardized positional scores by adjusting the distribution parameter lambda (describing mainly the distribution width). In the new version, in order to make the rescaled distribution closer to standard, the mean of the distribution is also forced to a fixed value. As a result, positional scores are more compatible with the gap penalties that were empirically optimized for the standard substitution matrices (e.g. BLOSUM 62). The optimal alignments on average become longer and cover similarity regions better without compromising the overall alignment accuracy.

(iii) *Improved speed*. Several algorithmic modifications, as well as a general code optimization, lead to an order of magnitude improvement in computational speed over the original version. The resulting computational efficiency is now comparable to that of the fastest profile-profile methods (12,15), with a typical search taking a few minutes on one processor. This time period may increase when the server is heavily loaded or when the user requires generation of the query profile by PSI-BLAST search, which may take longer for queries with a large number of homologs in the sequence database.

(iv) *Flexible control of input options*. The server's front page (Figure 1A) allows the user to upload the query in several common alignment formats, choose the database and adjust search parameters and output options. The query MSA or single sequence can be either pasted in the input window or uploaded from a file. The available profile databases currently include PFAM (32), COG, KOG (33,34) and PSI-BLAST alignments produced from sequences with known 3D structure: chain representatives of the PDB database (35) and domain representatives of SCOP classification (36). The PDB representatives are full chains extracted from the whole set of available 3D structures (35), based on a 70% cutoff of sequence identity. The SCOP representatives are structural domains defined and classified by expert analysis into families, superfamilies, folds and classes (36). These representatives are based on 40% identity and are taken from the ASTRAL database (37). The PDB and ASTRAL sequences are used as queries for PSI-BLAST searches against NCBI nr database. The resulting MSAs of detected homologs are used to generate COMPASS profiles. To allow for the choice of different levels of sequence divergence within MSAs, the user can choose profiles corresponding to different numbers of PSI-BLAST iterations. PFAM (32), COG and KOG (33,34) databases include families of both known and unknown 3D structure, which cover protein sequence space more completely and provide alternative ways of family classification. These databases typically represent tighter sequence grouping, with more consideration of protein function, and clustering of orthologs from different genomes. PFAM profiles are generated by



COMPASS

COmparison of Multiple Protein sequence Alignments with assessment of Statistical Significance.
Similarity search with alignment (or sequence) as a query against a database of protein families.

Enter query as protein sequence or multiple alignment:

Or file name C:\DUF185.aln

Choose database SCOP40_iter5
Email address to receive the result (optional)
Job ID (optional) duf185

Input processing options
Run PSI-BLAST if input is: a sequence ● Yes ○ No an alignment ○ Yes ● No
PSI-BLAST options: Iterations 5 Evalue 1e-3 Coverage 20 Identity 0
Gap fraction threshold 0.5

Search options
Gap opening penalty 10 Gap extention penalty 1 Matrix BLOSUM62
Effective length of the database ● Real Size ○ User Defined: 0

Output formatting options
Expect 1.0 Significance threshold 0.005 Display up to 100 hits
Top sequences to show in query 1 in subject 1
Show consensus sequences ● Yes ○ No Show fully gapped positions ○ Yes ● No
Width of alignment segment 60

Documentation

Downloads

COMPASS

```
COMPASS 2.42
Reference:
Ruslan I. Sadreyev and Nick V. Grishin (2003) COMPASS: a tool for comparison
of multiple protein alignments with assessment of statistical significance.
J Mol. Biol. 326(1):317-36

Job ID: duf185

Query = DUF185.aln
length=671; filtered_length=328; Neff=13.296
(threshold of effective gap content in columns: 0.5)

Database: SCOP40_iter5
Effective database length used: 1288522

• • •
                Hits with E-value BETTER than threshold:

Profiles producing significant alignments:                      Score   E-value
d1jqea_    c.66.1.19  Histamine methyltransferase                 139     4.28e-04
d1i4wa_    c.66.1.24  Transcription factor sc-mtTFB               111     4.03e-03


 Hits with E-value WORSE than threshold (may be not biologically meaningful):

d1nw3a_    c.66.1.31  Catalytic, N-terminal domain of histone m  120     7.34e-03
d1u2za_    c.66.1.31  Catalytic, N-terminal domain of histone m  117     1.01e-02
d1f3la_    c.66.1.6  Arginine methyltransferase, HMT1            113     1.47e-02
d1qama_    c.66.1.24  rRNA adenine dimethylase                   112     2.39e-02
d1af7_2    c.66.1.8  Chemotaxis receptor methyltransferase CheR  106     2.45e-02
d1inla_    c.66.1.17  Spermidine synthase                        111     2.87e-02
d1qyra_    c.66.1.24  High level kasugamycin resistance protein  109     3.14e-02
d1p91a_    c.66.1.33  rRNA methyltransferase RlmA                110     3.52e-02
d1nkva_    c.66.1.21  Hypothetical Protein YjhP                  109     4.58e-02
• • •
```

```
Subject = d1i4wa_    c.66.1.24  Transcription factor sc-mtTFB
length=322      filtered_length=314     Neff=17.121
Smith-Waterman score = 111      Evalue = 4.03e-03

Q9ZEA0_RICPR/60-305    10      WQRIGNP-KSLSLVELGPGRGLLMRDLLRT-AKLVPEFYKALSIKLIEINKNFIAHQKSN  68
CONSENSUS_1            14      WKLLGRPYADLSLVELGAGRGTLAADLLRALEKLGPELFERLSYHLVELSPALRARQRKT  74
                               +++ +R  ++ +++E+G+G G L++ LL        E++++ +++++E++P+L+++ R++
CONSENSUS_2            35      LLAGERVGPGDTVLEIGPGPGALTRALL========ERAGAARVVAIEIDPRLIELLRER  87
d1i4wa_                35      LTKTYKHPEELKVLDLYPGVGIQSAIFY========NKYCPRQYSLLEKRSSLYKFLNAK  87

Q9ZEA0_RICPR/60-305    69      LQ--------------DINLPIK------HLA---F-IEEIPQ-K-PTIIITNEFFDTM  101
CONSENSUS_1            75      LAPKLAPVGHDDSHVEIGLDGLVRDKGVIIWLDLESSLLEELPDDGVPGLLLANEFFDAL  134
                               LA +  ++ ++++++++ L++L++DK++++     ++L++  +++G P  +++N +  ++
CONSENSUS_2            88      LAGD=NLEVIHGDALKLDLLLLLIDKDAVL=====AGLILLADLPGDPLKVVGN=LPYNI  140
d1i4wa_                88      FEGS=PLQILKRDPYDWSTYSNLIDEERIF=====VPEVQSSDHINDKFLTVAN=VTGEG  140

Q9ZEA0_RICPR/60-305    102     P-IKQY----------IKVKELWYER-IFL  119
CONSENSUS_1            135     PHVHRFESGQEHGSLLVKTDGGWRERYVGV  164
                               +   +F+ ++++++ +   + G++++++ +
CONSENSUS_2            141     STPLLFKLLEALAPRRLLFRYGLVRLVLML  170
d1i4wa_                141     SEGLIMQWLSCIGNKNWLYRFGKVKMLLWM  170
```

Figure 1. (A) Front page of the COMPASS server. The main section allows the user to submit the query (by pasting in the window or by specifying the file), to choose the search database, and (if needed) to enter the email address to receive the results. The section of input processing options allows the user to choose whether a PSI-BLAST run is needed to enrich the query profile with additional sequence homologs and to define the parameters of profile construction. The section of search options can be used to adjust the main parameters of the search. The section of output options allows for flexible formatting of the search results. A brief explanation of each option is available by clicking on the option's name. Additional sections include the links to more detailed documentation and to the FTP page with standalone COMPASS package. **(B)** Search results for uncharacterized PFAM DUF185 as a query, supporting the structure and function prediction for this family. The list of hits among SCOP domains consistently includes members of the same superfamily of *S*-adenosyl-L-methionine-dependent methyltransferases (SAM-Mtases) (c.66.1). **(C)** Example of profile-profile alignment. The header includes brief information about the hit: database identifier, protein description, full length of the MSA ('length'), the length of the profile after purging positions with high gap content ('filtered length'), effective number of sequences as a characteristic of sequence divergence within MSA ('Neff'), followed by COMPASS score and E-value. In this example, the top and consensus sequences for compared profiles are displayed. Position matches with positive scores are marked by '+', identical residues in the two consensus sequences are marked by the residue symbol. Invariant glutamates of Motifs I and II (39) involved in ligand binding are marked with red dots, glycine-rich motif is circled. **D:** A recently solved structure for a member of DUF185 family (PDB ID 1zkd) confirms our prediction. Side chains of the invariant glutamate residues are shown in red, glycine-rich loop is circled.

COMPASS from full family alignments provided by PFAM. COG and KOG profiles are generated from MSAs produced from the database sequences by MUSCLE (38). The profile databases are regularly updated when new versions of original databases are available.

In order to gain more confidence in detected similarities and to find the best search conditions for a specific query, tuning the parameters controlling the generation of profiles and the construction of profile-profile alignments is advisable. The user can modify several such parameters. First, the input MSA (or sequence) can be used as a query for PSI-BLAST search, in order to produce a more diverse MSA of this family. The user can adjust the maximal number of iterations, as well as the requirements for a detected homolog to be included in the alignment: maximal E-value, minimal coverage of the query and minimal sequence identity to the query. Second, 'Gap fraction threshold' allows the user to control the maximal content of gaps in the MSA columns included in the COMPASS profile. If a column contains too many gaps, it is disregarded in the process of profile comparison, and shown in the final output as lower-case letters for residues and dots for gaps. The default value of this parameter is 0.5.

In the construction of profile-profile alignments, 'Gap penalties' are score penalties for opening and extending a new gap. 'Effective length of the database' is the parameter used in the calculation of E-values for the profile-profile alignments. For a given optimal alignment score, there is roughly a linear dependence of E-value on the assumed database length. 'Matrix' is a substitution matrix of the user's choice, BLOSUM62 by default. As described above, the choice of the matrix affects the rescaling of scores between individual profile positions that are used in the construction of the profile-profile alignment. Changing the scale of the positional scores would (i) make gap insertion more or less likely, affecting the resulting alignments, and (ii) change the optimal alignment scores and E-values.

Among the output formatting options, many are similar to those of PSI-BLAST. 'Expect' and 'significance threshold' are, respectively, the E-value cutoffs for the hit to be included in the output and to be considered significant. The hits outside the significance threshold are shown as potentially not meaningful. The user can also limit the total number of hits to display ('Display up to'). Some output options are specific to profile-profile comparison. For example, the displayed profile-profile alignments can include different numbers of top sequences from the input MSAs ('Top sequences to show'), as well as consensus sequences ('Show consensus sequences'). Brief help sections are provided for every adjustable parameter, as well as a link to more detailed documentation (Figure 1A).

(v) *User-friendly output*. The general structure of the output is similar to that of PSI-BLAST: the list of top hits is sorted by E-value and split into those below and above the significance threshold, followed by optimal profile-profile alignments with brief information about each hit. However, there are several significant differences, mainly in the format of alignments. The user can display the consensus sequences of profiles, as well as multiple top sequences from the input MSA. The number of top sequences displayed can range from zero (to show consensus only) to all sequences of the MSA. The complete query MSA is retrieved by clicking on the consensus link. Another feature for fast and convenient analysis is links to the original databases, which provide immediate access to information available for detected protein families.

Examples of remote similarity detection

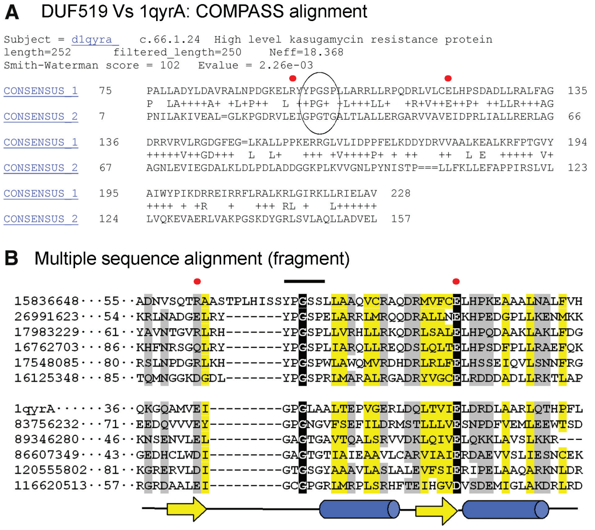
As an illustration, we describe the detection of distant sequence similarities that lead to fold predictions for two uncharacterized PFAM families annotated as 'DUF' (domain of unknown function). First, the COMPASS server detects homology between DUF185 (corresponding to COG1565 of the COG database) and SCOP domains of the *S*-adenosyl-*L*-methionine-dependent methyltransferase (SAM-Mtase) fold. Using the full DUF185 (PFAM 19.0) alignment as a query, with the default input parameters (Figure 1A), the server returns a list of hits that consistently belong to the same SCOP superfamily (c.66.1), both above and below the E-value cutoff (Figure 1B). In this list, each line consists of four fields: the identifier in the original database (implemented as a link to the database), a brief description of the protein, the COMPASS score and the corresponding E-value.

The next section of the output includes profile-profile alignments between the query and the hits. Each alignment is accompanied by a header with a brief information about the hit. Unlike the PSI-BLAST format, the alignments can include different numbers of top sequences from input MSAs and/or consensus sequences. Figure 1C shows an example of such an alignment, with a single top sequence and consensus displayed for each profile. To distinguish the gaps introduced by COMPASS from the gaps that already occur in the input alignments, the former are shown as equal signs (=). The alignment in Figure 1C includes the region of similarity between the query (profile for DUF185) and a homologous profile based on the PSI-BLAST alignment for structural domain 1i4wA. In addition to similar patterns of hydrophobicity and small residues, DUF185 shows a strong conservation of SAM-Mtase signature motifs [reviewed in (39)]. The SAM-binding loop GxGxG (circled) and conserved acidic residue in the preceding β-strand (marked with a red dot) are parts of Motif I, whereas the invariant glutamate at the end of the next β-strand (marked with a red dot) is a part of Motif II (39).

This previously published prediction had been difficult to produce by PSI-BLAST, even for an expert user (24). However, it was more recently confirmed by the solved structure of a DUF185 member. This structure (PDB ID 1zkd, Northeast Structural Genomics Consortium) has been neither functionally annotated nor classified by SCOP or CATH, but possesses typical features of the SAM-Mtase fold (Figure 1D). The core of the domain contains a mixed β-sheet of seven β-strands surrounded by two sheets of α-helices. The strand order is 3214576; with strand 7 (colored red) anti-parallel to the rest and forming a characteristic methyltransferase β-hairpin with strand 6 (colored orange). In this domain, the β-hairpin contains an additional α-helical insert (orange helices). The presence of a glycine-rich loop (circled) and other signature motifs, including glutamates marked in Figure 1C (side chains shown in red), suggest that this domain is a functional methyltransferase.

The second prediction originates from searching with RrnaAD methylase family as a query. This search reveals a newly identified similarity to a PFAM family of mainly hypothetical bacterial proteins with unknown structure and function, DUF519 (corresponding to COG2961 in the COG database). Thus, we suggest that DUF519/COG2961 proteins also possess the structural SAM-Mtase fold. This hypothesis is supported by the results of a search with the PFAM 19.0 DUF519 alignment as a query against the database of SCOP profiles (PSI-BLAST iteration 3). Homologs detected above the significance threshold, as well as multiple hits below the threshold, consistently belong to the SAM-Mtase fold.

Figure 2A shows the COMPASS alignment between DUF519 and the detected homolog, a domain of the SAM-Mtase fold (PDB ID 1qyrA). This domain (not shown) possesses typical features of the fold and is similar to the structure shown in Figure 1D. Figure 2A shows the COMPASS alignment including the signature motifs of SAM-Mtases. Figure 2B shows the MSA of representatives from both families that covers SAM-Mtase Motifs I and II (39). In DUF519, this region includes the invariant glutamate aligned to a ligand-binding glutamate of SAM-Mtases (E95 in the top sequence, marked with red dot), the characteristic location of conserved small residues in the SAM-binding loop (marked with a line) and a similar hydrophobicity pattern. Secondary structure prediction for this part of DUF519 is also consistent with the secondary structure of the SAM-Mtase fold. This prediction is additionally supported by other tools, e.g. by (i) significant scores for the similarity with the SCOP SAM-Mtase domains produced by FFAS03 server (14); and (ii) the results of multiple iterations of PSI-BLAST search in a sequence database with a family representative as a query. After four iterations, PSI-BLAST detects the similarity between a DUF519 sequence Q9PHA1_XYLFA (gi|15836648, residues 32-291) and two proteins of known structure possessing the SAM-Mtase fold (PDB IDs 2ift and 2fpo).


A
DUF519 Vs 1qyrA: COMPASS alignment
B
Multiple sequence alignment (fragment)

```
Subject = d1qyra_   c.66.1.24  High level kasugamycin resistance protein
length=252     filtered_length=250    Neff=18.368
Smith-Waterman score = 102   Evalue = 2.26e-03

CONSENSUS_1   75     PALLADYLDAVRALNPDGKELRYYPGSPLLARRLLRPQDRLVLCELHPSDADLLRALFAG  135
                     P  LA++++A+ +L+P++  L ++PG+ +L+++LL+  +R+V++E++P+ ++LLR+++AG
CONSENSUS_2   7      PNILAKIVEAL=GLKPGDRVLEIGPGTGALTLALLERGARVVAVEIDPRLIALLRERLAG  66

CONSENSUS_1   136    DRRVRVLRGDGFEG=LKALLPPKERRGLVLIDPPFELKDDYDRVVAALKEALKRFPTGVY  194
                     +++++V++GD+++  L  L+   +++ +V++++P+ + ++   ++  L E  +++++ V+
CONSENSUS_2   67     AGNLEVIEGDALKLDLPDLADDGGKPLKVVGNLPYNISTP===LLFKLLEFAPPIRSLVL  123

CONSENSUS_1   195    AIWYPIKDRREIRRFLRALKRLGIRKLLRIELAV  228
                     ++++ + +R   +   +++ RL +   L++++++
CONSENSUS_2   124    LVQKEVAERLVAKPGSKDYGRLSVLAQLLADVEL  157
```

```
15836648···55··ADNVSQTRAASTPLHISSYPGSSLLAAQVCRAQDRMVFCELHPKEAAALNALFVH
26991623···54··KRLNADGELRY-------YPGSPELARRLMRQQDRALLNEKHPEDGPLLKENMKK
17983229···61··YAVNTGVRLRH-------YPGSPLLVRHLLRKQDRLSALELHPQDAAKLAKLFDG
16762703···86··KHFNRSGQLRY-------YPGSPLIARQLLREQDSLQLTELHPSDFPLLRAEFQK
17548085···80··RSLNPDGRLKH-------YPGSPWLAWQMVRDHDRLRLFELHSSEIQVLSNNFRG
16125348···85··TQMNGGKDGDL-------YPGSPRLMARALRGADRYVGCELRDDDADLLRKTLAP

1qyrA······36··QKGQAMVEI---------GPGLAALTEPVGERLDQLTVIELDRDLAARLQTHPFL
83756232···71··EEDQVVVEY---------GPGNGVFSEFILDRMSTLLLVESNPDFVEMLEEWTSD
89346280···46··KNSENVLEL---------GAGTGAVTQALSRVVDKLQIVELQKKLAVSLKKR---
86607349···43··GEDHCLWDI---------GAGTGTIAIEAAVLCARVIAIERDAEVVSLIESNCEK
120555802··81··KGRERVLDI---------GTGSGYAAAVLASLALEVFSIERIPELAAQARKNLDR
116620513··57··RGRDAALEI---------GCGPGRLMRPLSRHFTEIHGVDVSDEMIGLAKDRLRD
```

Figure 2. Search results for PFAM DUF519 suggest that this family possesses the structural fold of SAM-Mtases. (**A**) DUF519 is used as a query for the COMPASS search against the databases of PSI-BLAST alignments (iteration 3) for SCOP representatives. The COMPASS alignment between the query and the detected homolog (domain 1qyrA) includes characteristic motifs of the SAM-Mtase superfamily. In this example, only consensus sequences are displayed. Positions corresponding to the conserved acidic residues of Motifs I and II (39) are marked with red dots. The region of the SAM-binding loop is circled. (**B**) Multiple alignment including representatives from DUF519 (top) and 1qyrA homologs (bottom). Sequences are denoted by NCBI GI numbers. Positions corresponding to conserved acidic residues of SAM-Mtase are marked with red dots. The region of the ligand-binding loop is marked with a line. Invariant residues are boxed in black. Uncharged residues (all amino acids except D, E, K, R) in mostly hydrophobic sites are highlighted in yellow; non-hydrophobic residues (all amino acids except W, F, Y, M, L, I, V) at mostly hydrophilic sites are highlighted in light gray. The secondary structure of 1qyrA is shown below the alignment, with α-helices and β-strands displayed as cylinders and arrows, respectively.

ACKNOWLEDGEMENTS

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources. We would like to thank Lisa Kinch and James Wrabl for discussions and critical reading of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
2. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
3. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Karplus,K., Barrett,C., Cline,M., Diekhans,M., Grate,L. and Hughey,R. (1999) Predicting protein structure using only sequence information. *Proteins*, **37,** (Suppl. 3), 121–125.
5. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, **53**(Suppl. 6), 491–496.
6. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
7. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
8. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
9. Sadreyev,R.I. and Grishin,N.V. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
10. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–581.
11. Edgar,R.C. and Sjolander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309–1318.
12. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
13. Frenkel-Morgenstern,M., Singer,A., Bronfeld,H. and Pietrokovski,S. (2005) One-Block CYRCA: an automated procedure for identifying multiple-block alignments from single block queries. *Nucleic Acids Res.*, **33**, W281–W283.

14. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. and Godzik,A. (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.*, **33**, W284–W288.
15. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
16. Soding,J., Remmert,M., Biegert,A. and Lupas,A.N. (2006) HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Res.*, **34**, W374–W378.
17. Ohlson,T., Wallner,B. and Elofsson,A. (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins*, **57**, 188–197.
18. Wang,G. and Dunbrack,R.L.Jr. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci.*, **13**, 1612–1626.
19. Chivian,D., Kim,D.E., Malmstrom,L., Schonbrun,J., Rohl,C.A. and Baker,D. (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, **61**(Suppl. 7), 157–166.
20. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
21. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
22. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
23. Zhou,H. and Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*, **61**(Suppl. 7), 152–156.
24. Sadreyev,R.I., Baker,D. and Grishin,N.V. (2003) Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci.*, **12**, 2262–2272.
25. Birtle,Z. and Ponting,C.P. (2006) Meisetz and the birth of the KRAB motif. *Bioinformatics*, **22**, 2841–2845.
26. Kim,B.H., Sadreyev,R. and Grishin,N.V. (2005) COG4849 is a novel family of nucleotidyltransferases. *J. Mol. Recognit.*, **18**, 422–425.
27. Theobald,D.L., Cervantes,R.B., Lundblad,V. and Wuttke,D.S. (2003) Homology among telomeric end-protection proteins. *Structure*, **11**, 1049–1050.
28. Theobald,D.L. and Wuttke,D.S. (2004) Prediction of multiple tandem OB-fold domains in telomere end-binding proteins Pot1 and Cdc13. *Structure*, **12**, 1877–1879.
29. Theobald,D.L. and Wuttke,D.S. (2005) Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *J. Mol. Biol.*, **354**, 722–737.
30. Wels,M., Francke,C., Kerkhoven,R., Kleerebezem,M. and Siezen,R.J. (2006) Predicting cis-acting elements of Lactobacillus plantarum by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res.*, **34**, 1947–1958.
31. Winter,E.E. and Ponting,C.P. (2005) Mammalian BEX, WEX and GASP genes: coding and non-coding chimaerism sustained by gene conversion events. *BMC Evol. Biol.*, **5**, 54.
32. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
33. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
34. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
35. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
36. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
37. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
38. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
39. Schubert,H.L., Blumenthal,R.M. and Cheng,X. (2003) Many paths to methyltransfer: a chronicle of convergence. *Trends Biochem. Sci.*, **28**, 329–335.